

A factor model approach for the joint segmentation with between-series correlation

X. Collilieux^{a,b}, E. Lebarbier^{c,d,*}, S. Robin^{d,c}

- ^(a) IGN LAREG, Univ Paris Diderot, Paris, France
- ^(b) Observatoire de Paris, SYRTE, CNRS, UPMC, Paris, France
- ^(c) AgroParisTech, UMR 518 MIA, Paris, France
- ^(d) INRA, UMR 518 MIA, Paris, France
- ^(*) corresponding author: lebarbie@agroparistech.fr

Abstract. We consider the segmentation of set of correlated time-series due e.g. to some spatial structure. We propose to model the between-series dependency with a factor model. This modeling allows us to use the dynamic programming algorithm for the inference of the breakpoints, which remains the most efficient strategy. We also propose a model selection procedure to determine both the number of breakpoints and the number of factors. The performance of our proposed procedure is assessed through simulation experiments. An application to geodetic data is presented.

Keywords. Dynamic programming; EM algorithm; Factor model; Segmentation; Model selection.

1 Introduction

The objective of segmentation methods is to detect abrupt changes in the distribution of a signal. Such segmentation problems arise in many areas: in biology for the detection of chromosomal aberrations ([9, 5]), in climatology for the detection of changes due to instrumental changes ([2, 6]) or in geodesy for the detection of changes in GPS location series either due to instrumental or to environmental changes as earth's crust shifts [12].

In many cases, we have at hand multiple series (several patients or GPS receivers) and it is likely that some dependency exist between them. Probe effect in the genomic context [7] or spatial correlation in climatology and geodesy context (due to their spatial organization) typically results in such a between-series dependency. Dealing with multiple series requires to account for these correlations in order to avoid false breakpoint detection (see [8]).

We consider here the detection of breakpoints in a set of time-series. We are typically interested in the case where this data consists of series of measurements observed along time in different locations. Each series is supposed to be affected by changes at series-specific breakpoints and the signals observed at each location are supposed to be correlated from one series to another. One of the difficulty here is to propose a modeling for which an efficient estimation algorithm can be derived. Indeed, the inference of segmentation models requires to search over the space of all possible segmentations, which is prohibitive in terms of computational time, when performed in a naive way. The Dynamic Programming (DP) strategy is the only one that gives the exact solution in a fast way but only applies when the contrast to be optimized is additive with respect to the segments (see [1, 2, 9]). In presence of dependency, the contrast (e.g. the log-likelihood) is generally not additive. In some

sense, our strategy consists in 'removing' the dependency so that, at a certain step of the estimation algorithm, DP can be applied to transformed data.

A similar setting is considered in [7] where a variance component model is used to account for the dependency between the series. This case consists in the simplest correlation form we can have, that is a uniform correlation between all series. Our purpose here is to broad the set of possible dependency structure that the modeling can account for. The factor model provides a convenient and efficient way to model covariance matrices, possibly limiting the number of parameters. It can be viewed as a generalization of variance components models, where the components are unknown and need to be estimated, together with the associated variances. It is based on the spectral decomposition of the covariance matrix and has been successfully applied in situations where very little is known about the correlation structure (see e.g. [3]). The inference of factor models can be achieved via an EM algorithm.

We present here a general model for correlated Gaussian series, based on a factor model for the covariance matrix. We show that some by-product of the EM algorithm can be used to remove the dependency between the series. This allows us to combine the EM and DP algorithms together. In terms of model selection, we propose a heuristic procedure combining two BIC criteria: the classical BIC for the choice of the number of factors and the modified BIC criterion derived by [13] in the context of segmentation for the choice of the number of segments.

The article is organized as follows. In Section 2, we present the proposed model for multiple series which is a segmentation model combined with a factor model. The EM algorithm proposed to obtain the maximum likelihood estimates is described in Section 3. In Section 4, we introduce a model selection procedure for both the number of segments and the number of factors. A simulation study is performed in Section 5 and in Section 6 we apply our method to geodetic data.

2 Model

We consider M series with n points each. We note y_{tm} the observed signal of series m at time t . The total number of observations is $N = nM$. The data are gathered in a matrix \mathbf{Y} with dimension $[n \times M]$. For a given matrix \mathbf{A} , we denote by A_t its t -th row and by A^m its m -th column. Thus the column Y^m represents whole series m , while the row Y_t contains the observations at time t in all the series.

Breakpoints and segments We consider here that each series has its own breakpoints: the mean of the series $Y^m = (Y_{tm})_{t=1..n}$ is subject to $K_m - 1$ specific abrupt changes at breakpoints $(t_k^m)_{k=0..K_m}$ (with convention $t_0^m = 0$ and $t_{K_m}^m = n$) and is constant between two breakpoints within the interval $I_k^m = [t_{k-1}^m + 1; t_k^m]$. In the following we denote by $K = \sum_{m=1}^M K_m$ the total number of segments and $n_k^m = t_k^m - t_{k-1}^m$ the length of segment k from series m ($k = 1, \dots, K_m$). The segmentation model is written as follows:

$$Y_{tm} = \mu_{km} + F_{tm} \quad \forall t \in I_k^m, \quad (1)$$

where the M -dimensional error vectors $\{F_t\}_{t=1..n}$ are supposed to be i.i.d. centered Gaussian and with covariance matrix Σ .

Correlations between series We model the dependency between the series via a factor model. More precisely, we assume that the dependency across series at each time t is captured by some latent factor Z_t , with dimension Q , which affects all residual terms through a matrix \mathbf{B} :

$$F_{tm} = \sum_{q=1}^Q Z_{tq} b_{qm} + E_{tm} \quad \forall t, m \quad \Longleftrightarrow \quad F_t = Z_t \mathbf{B}' + E_t \quad \forall t$$

where $\mathbf{B} = (b_{mq})$ is a fixed $[M \times Q]$ matrix and where the random vectors $Z_t = (Z_{tq})$ are i.i.d., centered, Gaussian with covariance matrix \mathbf{I}_Q and independent from the random vectors $E_t = (E_{tm})$, which are i.i.d. centered Gaussian with diagonal covariance matrix Ψ . Note that assumption $\mathbb{V}(Z_t) = \mathbf{I}_Q$ is necessary for identifiability reasons.

As a result, the covariance matrix Σ of the F_t writes

$$\Sigma = \mathbf{B}\mathbf{B}' + \Psi. \quad (2)$$

In this decomposition, $\mathbf{B}\mathbf{B}'$ refers to the shared variance and Ψ to the specific one.

An important feature of the factor model is that any M -dimensional covariance matrix can be recovered, provided that $Q = M - 1$. As an example, the classical spatial correlation structure $(\Sigma)_{m,m'} \propto e^{-d(m,m')}$ where $d(m,m')$ stands for the distance between the locations of series m and m' can be rewritten under the form (2). Another interest of the factor model is that a small value Q provides regularization of the covariance matrix.

Model With the previous decomposition of the variability, the model (1) can be rewritten as a mixed linear model:

$$Y_{tm} = \mu_{km} + \sum_{q=1}^Q Z_{tq} b_{qm} + E_{tm} \quad \forall t \in I_k^m.$$

This linear model can be written with the following matrix form:

$$\mathbf{Y} = \mathbf{T}\boldsymbol{\mu} + \mathbf{Z}\mathbf{B}' + \mathbf{E}, \quad (3)$$

where \mathbf{Y} , with size $[n \times M]$, stands for the observed data, \mathbf{T} is the incidence matrix of breakpoints with size $[n \times K]$, $T^m = \text{Bloc} \left[\mathbb{1}_{n_{K_m}^m} \right]$, and $\mathbf{T} = [T^1 \ T^2 \ \dots \ T^M]$, $\boldsymbol{\mu}$ is the means with size $[K \times M]$ (and μ_k^m the mean of the segment k for series m) such that $\mu^m = \text{Bloc} [\mu_{K_m}^m]$, and $\boldsymbol{\mu} = [\mu^1 \ \mu^2 \ \dots \ \mu^M]$, \mathbf{Z} with size $[n \times Q]$ and \mathbf{B} with size $[M \times Q]$, and \mathbf{E} with size $[n \times M]$.

The main difference between this model and a classical mixed linear model is that both the incidence matrix \mathbf{T} and the factor matrix \mathbf{B} are unknown.

Parametrization This detour through the factor model is motivated by an algorithmic issue. Indeed, dealing with a covariance matrix which is not diagonal as for example $(\Sigma)_{m,m'} \propto e^{-d(m,m')}$ hampers the direct use of the DP algorithm (used for the inference of the breakpoints). As shown in the next section, the reparametrization of Σ as (2) where Ψ is diagonal allows us to use it.

To estimate all the parameters of the model, we proceed as classically in two steps: we estimate the parameters $\phi = (\mathbf{T}, \boldsymbol{\mu}, \sigma^2, \mathbf{B})$, the number of factors Q and the number of segments K being fixed (Section 3), then we use a model selection strategy to choose Q and K (Section 4).

3 Estimation using the EM algorithm

To estimate the set of parameters ϕ , we consider the maximum-likelihood procedure. This can be done via an EM algorithm, considering that \mathbf{Z} represents the missing (or latent) variables [10]. This algorithm is now classical for the inference in mixed linear model (see for example [11]). We remind that the key quantity in the EM algorithm is the conditional expectation, $Q(\phi; \phi^{(h)}) = \mathbb{E}_{\phi^{(h)}} \{\log \mathcal{L}(\mathbf{Y}|\mathbf{Z}; \phi) | \mathbf{Y}\}$ where the complete-data log-likelihood is

$$\log \mathcal{L}(\mathbf{Y}, \mathbf{Z}; \phi) = \log \mathcal{L}(\mathbf{Y}|\mathbf{Z}; \phi) + \log \mathcal{L}(\mathbf{Z}).$$

Since the distribution of \mathbf{Z} does not depend on the parameters ϕ , only the first term will be considered which writes

$$-2 \log \mathcal{L}(\mathbf{Y}|\mathbf{Z}; \phi) = N \log(2\pi) + n \log(|\Psi|) + \sum_{t=1}^n \|Y_t - \mu_{k(t)} - Z_t \mathbf{B}'\|_{\Psi^{-1}}^2,$$

then its conditional expectation $Q(\phi; \phi^{(h)})$ satisfies

$$\begin{aligned} -2Q(\phi; \phi^{(h)}) &= N \log(2\pi) + n \log(|\Psi|) \\ &\quad + \sum_{t=1}^n \left[\|Y_t - \mu_{k(t)} - \hat{\mathbf{Z}}_t^{(h)} \mathbf{B}'\|_{\Psi^{-1}}^2 + \text{Tr} \left(\mathbf{B}' \Psi^{-1} \mathbf{B} \mathbf{W}_t^{(h)} \right) \right], \end{aligned}$$

where $\phi^{(h)}$ is the value of ϕ at iteration (h) , $\mathbb{E}_\phi\{\cdot\}$ is the expectation calculated with ϕ as the parameter value and $\mathbb{V}_\phi\{\cdot\}$ the corresponding variance, $\hat{\mathbf{Z}}_t^{(h)} = \mathbb{E}_{\phi^{(h)}}\{\mathbf{Z}_t|\mathbf{Y}\}$, and $\mathbf{W}_t^{(h)} = \mathbb{V}_{\phi^{(h)}}\{\mathbf{Z}_t|\mathbf{Y}\}$. $\text{Tr}(A)$ stands for the trace of matrix A and $|A|$ for its determinant.

The EM algorithm is an iterative algorithm which consists in two steps (E-step and M-step) at each iteration. At iteration $(h+1)$, we have

E-step This step consists in the calculation of the conditional expectation $Q(\phi; \phi^{(h)})$ which requires the conditional moments $\hat{\mathbf{Z}}$ and \mathbf{W} . We get

$$\begin{aligned} \hat{\mathbf{Z}}_t^{(h+1)} &= \tilde{Y}_t^{(h)} \mathbf{B}^{(h)} \mathbf{W}_t^{(h)} / \sigma^{2,(h)}, \\ \mathbf{W}_t^{(h+1)} &= \left(I_M + \mathbf{B}'^{(h)} \mathbf{B}^{(h)} / \sigma^{2,(h)} \right)^{-1}, \end{aligned}$$

where $\tilde{Y}_t^{(h)} = Y_t - \mu_{k(t)}^{(h)}$.

M-step This step consists in the estimation of the parameters by maximizing the obtained conditional expectation. We get

- Estimation of \mathbf{B} :

$$\mathbf{B}^{(h+1)} = \sum_{t=1}^n (Y_t - \mu_{k(t)}^{(h)})' \hat{\mathbf{Z}}_t^{(h+1)} \left[\sum_{t=1}^n (\hat{\mathbf{Z}}_t^{(h+1)} \hat{\mathbf{Z}}_t^{(h+1)} + \mathbf{W}_t^{(h+1)}) \right]^{-1}.$$

- Estimation of residual covariance:

$$\Psi^{(h+1)} = \arg \max_{\Psi} Q_0 = \frac{1}{n} \sum_{t=1}^n (Y_t - \mu_{k(t)}^{(h)})' E_t^{(h)},$$

where $E_t^{(h)} = Y_t - \mu_{k(t)}^{(h)} - \hat{\mathbf{Z}}_t^{(h+1)} \mathbf{B}'^{(h+1)}$. In the case where $\Psi = \sigma^2 \mathbf{I}_M$, as we consider in the simulation study and the application, we get

$$\sigma^{2,(h+1)} = \frac{1}{N} \sum_{t=1}^n \left[E_t^{(h)} E_t'^{(h)} + \text{Tr} \left(\mathbf{B}'^{(h)} \mathbf{B}^{(h)} \mathbf{W}_t^{(h+1)} \right) \right]$$

and $\Psi^{(h+1)} = \sigma^{2,(h+1)} \mathbf{I}_M$.

- Estimation of the segmentation parameters $\mathbf{T}\boldsymbol{\mu}$:

$$\begin{aligned}\left\{\mathbf{T}^{(h+1)}, \boldsymbol{\mu}^{(h+1)}\right\} &= \arg \min_{\mathbf{T}, \boldsymbol{\mu}} \sum_{t=1}^n \|Y_t - (\mathbf{T}\boldsymbol{\mu})_t - \widehat{\mathbf{Z}}_t^{(h+1)} \mathbf{B}'^{(h+1)}\|_{(\boldsymbol{\Psi}^{(h+1)})^{-1}}^2, \\ &= \arg \min_{\mathbf{T}, \boldsymbol{\mu}} \sum_{t=1}^n \|\check{Y}_t - (\mathbf{T}\boldsymbol{\mu})_t\|_{(\boldsymbol{\Psi}^{(h+1)})^{-1}}^2,\end{aligned}$$

where $\check{Y}_t = Y_t - \widehat{\mathbf{Z}}_t^{(h+1)} \mathbf{B}'^{(h+1)}$. This last term can be viewed as a correction to remove dependency between the series. $\boldsymbol{\Psi}$ being diagonal, the DP algorithm can be used to obtain the segmentation parameter. In particular, we use the two-stages DP proposed by [7, 8] which is a fast version of DP dedicated to the joint segmentation of multiple series. In the case where $\boldsymbol{\Psi} = \sigma^2 \mathbf{I}_M$, the homoskedastic noise case, the quantity to be minimized is a residual sum of squares. In the heteroskedastic noise case, the considered sum of squares is its weighted version.

4 Model selection

Both the number of factors Q and the number of segments K should be estimated. This joint model selection issue is not classical and difficulty arises due to the different nature of the parameters at hand. Indeed the likelihood function is continuous with respect to the loadings \mathbf{B} of the latent vectors, so the classical framework for the BIC approximation holds. This is not true for the segmentation parameters and a specific BIC approximation needs to be derived, as observed by [13]. Furthermore, the two resulting criteria do not share the same form, so they can not be easily combined into a single criterion. We propose here a two-step heuristic to select these two parameters.

For a fixed number of segments K , we use the classical BIC criterion

$$BIC_K(Q) = 2\mathcal{L}(\widehat{\mathbf{T}\boldsymbol{\mu}}_K, \widehat{\boldsymbol{\Sigma}}_Q) - D_Q \log(n),$$

where $\boldsymbol{\Sigma}_Q$ is the covariance matrix for a given Q , $\mathbf{T}\boldsymbol{\mu}_K$ is the segmentation parameters for a given K , $\mathcal{L}(\widehat{\mathbf{T}\boldsymbol{\mu}}_K, \widehat{\boldsymbol{\Sigma}}_Q)$ is the log-likelihood calculated at its maximum for the model with K segments and Q factors, and D_Q is the number of parameters in a model with Q factors, $D_Q = Q(2M - Q + 1)/2 + 1$. Indeed for the variance components, according to the variance decomposition (cf equation (2)), the number of parameters are MQ for B and one for σ^2 . Moreover, with the orthogonality condition on B , only $MQ - Q(Q - 1)/2$ need to be estimated.

The number of factor Q is then selected as

$$\widehat{Q}_K = \arg \max_Q BIC_K(Q).$$

For each Q , we select the number of segments K with the modified BIC proposed by [13] and adapted to the joint segmentation of multiple series:

$$\begin{aligned}mBIC_Q(K) &= \left(\frac{K-M}{2}\right) \log\left(\frac{SS_{\text{all}}}{2}\right) + \left(\frac{N-K}{2} + 1\right) \log\left(1 + \frac{SS_{\text{bg}}(\widehat{t})}{SS_{\text{wg}}(\widehat{t})}\right) \\ &\quad + \log\left[\Gamma\left(\frac{N-K}{2} + 1\right)\right] - \frac{1}{2} \sum_{m=1}^M \sum_{k=1}^{k_m} \log \widehat{n}_k^m - (K-M) \log(N),\end{aligned}$$

with

$$\begin{aligned} SS_{\text{wg}}(\hat{t}) &= \sum_{t=1}^n (\mathbf{Y}_t - \hat{\mu}_{k(t)}) \hat{\Sigma}_Q^{-1} (\mathbf{Y}_t - \hat{\mu}_{k(t)})', \\ SS_{\text{all}} &= \sum_{t=1}^n (\mathbf{Y}_t - \bar{Y}) \hat{\Sigma}_Q^{-1} (\mathbf{Y}_t - \bar{Y})', \\ SS_{\text{bg}}(\hat{t}) &= SS_{\text{all}} - SS_{\text{wg}}(\hat{t}), \end{aligned}$$

\hat{n}_k^m is the length of segment k in series m ($\hat{n}_k^m = \hat{t}_k^m - \hat{t}_{k-1}^m$), and $\hat{\mu}_{k(t)}$ is a vector of size M with the component m is $\bar{y}_{mk} = (\hat{n}_k^m)^{-1} \sum_{t=\hat{t}_{k-1}^m+1}^{\hat{t}_k^m} y_m(t)$ if $t \in \hat{I}_k^m$.

The number of segments is the chosen as

$$\hat{K}_Q = \arg \max_K mBIC_Q(K).$$

We then propose the following heuristic: choose the best Q for each K , then select the best K among them:

$$\hat{Q}_K = \arg \max_Q BIC_K(Q), \quad \hat{K}_{\hat{Q}_K} = \arg \max_K mBIC_{\hat{Q}_K}(K) \quad \text{and} \quad \hat{Q} = \hat{Q}_{\hat{K}}.$$

5 Simulation study

In this section, we illustrate the importance of taking the dependency into account and we study the behavior of our model selection heuristic for K and Q and its impact on the estimation of all parameters.

5.1 Simulation design and quality criteria

Simulation design We consider different number of series $M \in \{5, 10\}$ with different lengths $n \in \{50, 100\}$. For each series Y^m , the number of breakpoints ($K-1$) is Poisson distributed with mean \bar{k} ($\bar{k} = 3$ for $n = 50$ and $\bar{k} = 5$ for $n = 100$) and their positions are uniformly distributed. The mean value within each segment alternates between 0 and values in $\{-2, -1, +1, +2\}$. We consider different residual standard deviations $\sigma \in \{.2, .5, 1\}$ and a spatial-type correlation between series: distances d are simulated as distances between Gaussian bivariate random vectors and $\Sigma = ((1-\alpha)\rho^d + \alpha I_M) \times \sigma^2$ with $\alpha = 0.2$ and $\rho \in \{0.2, 0.8\}$. The parameter ρ controls the intensity of the dependency between series. Each configuration is simulated 100 times.

Quality criteria To assess the quality of the estimation of the covariance matrix, we use the root mean squared distance between the true parameter and its estimate : $\text{RMSE}(\Sigma) = \left[M^{-2} \sum_{i,j=1}^Q (\hat{\Sigma}_{ij} - \Sigma_{ij})^2 \right]^{1/2}$. For the segmentation parameters, we are interested in the performance of the breakpoint positioning. To measure it, we consider both the proportion of erroneously detected breakpoints among detected breakpoints (false positive rate, FPR) and the proportion of detected true breakpoints among true detected breakpoints (true positive rate, TPR). A perfect segmentation results in null FPR and TPR equals to 1. For each configuration we consider the average of these criteria over the 100 simulations.

5.2 Results

Only the results with $M = 10$ and $n = 100$ are presented since the results for the other configurations lead to same conclusions. In this section, we use notations $*$ for the true parameters and est for the estimated ones on the graphs.

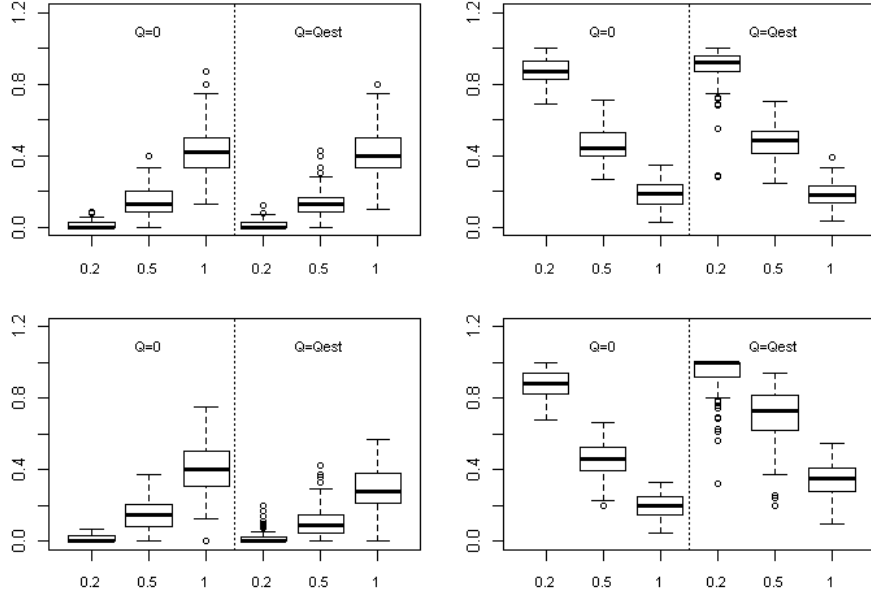


Figure 1: FPR on the left and TPR on the right for $\rho = 0.2$ (top) and $\rho = 0.8$ (bottom) using \hat{K} . We distinguish the cases $Q = \hat{Q}$ (denoted Qest on the graph) and $Q = 0$ (e.g. the segmentation only). x -axis: σ .

Accounting for the dependency Figure 1 compares the selected segmentation when the dependency is considered ($Q = \hat{Q}$) or not ($Q = 0$). Whatever the difficulty of the detection problem (different values of σ), accounting for the dependency increases the performance of the segmentation (smaller FPR and larger TPR). This is particularly marked when the dependency is high ($\rho = 0.8$, bottom of the figure).

Discussion on the selection of K Figure 2 (top) shows that, whatever the level of the dependency, when the noise is small ($\sigma = 0.2$), the selected number of segments is close to the true one and the breakpoints are well positioned (see Figure 2 (bottom)). When the detection problem gets difficult (σ large), the selection procedure tends to underestimate the number of segments leading to a better precision on the breakpoints positioning compared to the true number of segments (smaller FPR). This result was expected since in this case one may prefer to avoid the detection of false positive breakpoints, as generally observed in segmentation problems.

Discussion on the selection of Q The number of factors is strongly underestimated (see Table 1) (close to 1 in average for $\rho = 0.2$ and for the different values of σ , not shown) meaning that only few factors are necessary to capture the dependency structure. This underestimation does not alter the estimation of Σ (compared to the true number of factors) and increases the power of procedure in terms of breakpoint positioning (in terms of FPR and TPR). Moreover, the selected number of factors decreases slightly with the difficulty of the detection problem (with σ) leading to a decreasing precision of the estimation of Σ .

We observe the benefit of the factor model estimate in terms of regularization.

Confusion between \hat{K} and $\hat{\Sigma}$ As observed in Figure 2 (bottom), the true number of segments $K = K^*$ leads to numerous false positive breakpoints. In some sense, K^* is too large for the data at hand. The consequence of oversegmentation is that no factor are selected for $\sigma = 0.5, 1$, meaning that the dependency is captured by

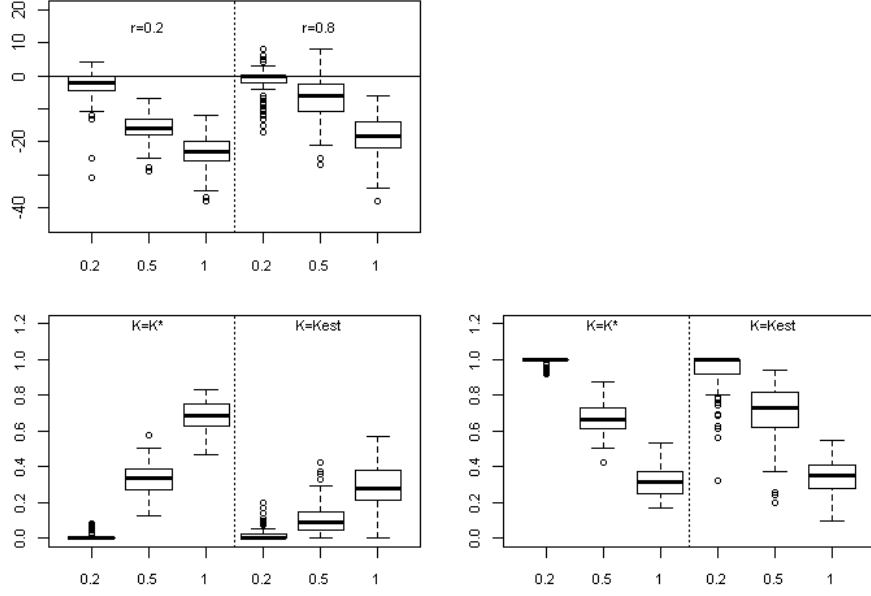


Figure 2: Top: $\hat{K} - K^*$ obtained with $Q = \hat{Q}$ for $\rho = (0.2, 0.8)$ (denoted r on the graph). Bottom: FPR (left) and TPR (right) with $Q = \hat{Q}$, $\rho = 0.8$ for $K = K^*$ and $K = \hat{K}$. x -axis: σ .

σ	(\hat{Q}, \hat{K})			(Q^*, \hat{K})		
	0.2	0.5	1	0.2	0.5	1
mean of Q	3.37	2.74	2.39	9		
RMSE(Σ)	0.005	0.034	0.119	0.0048	0.032	0.124
FPR	0.016	0.110	0.288	0.039	0.175	0.413
TPR	0.93	0.69	0.34	0.93	0.597	0.262

Table 1: Mean of Q , RMSE(Σ), FPR and TPR for $\rho = 0.8$ where Q^* is the true number of factors.

the segmentation (results not shown).

6 Application

Data description Scientific Permanent GNSS instruments continuously track electromagnetic signals from GPS satellites. Their data are generally computed by scientific, private or public services in near-real time or after getting a few days of observations to derive accurate coordinates. These coordinates are used to determine precise velocities of points located on the crust that constrain tectonic models and Earth’s crust/mantle parameters, to infer mass transfer in the fluid layers of the Earth or for positioning applications (terrestrial reference frame).

Coordinate time series from five GNSS stations have been used in this study. These stations, located in the Michigan and Ohio States in USA. They are encoded MPLE, ADRI, BAYR, BRIG and DEFI. Only the longitude component has been investigated here: from the February 2002 to June 2013, 3776 longitude coordinates are available per station. Because they are separated by less than 250 km, their coordinate time series show similarities related to common ground deformation and correlated processing errors. The predominant effect in this component is the tectonic motion of the North American plate, which is about -16 mm/yr. When making difference of coordinate series from close stations, we can hope that this effect is cancelled. Only residual differential motion as well as residual noise still remains in addition to sudden changes often related to equipment or environmental changes at one of the two stations.

Working with difference of series allows us to neglect the modelisation of the ground deformation (which can be complex, see [4]), that our model does not take into account, as in [2] for climate series. In this example, we use station MPLE as the reference series and form four time series from the four other slave series, denoted ADRI-MPLE, BAYR-MPLE, BRIG-MPLE and DEFI-MPLE.

Accounting for dependency For those series, a model where the dependency is considered is preferred according to the model selection criterion to the model where it is not the case ($Q = 0$). More precisely, our method selects one factor ($\hat{Q} = 1$) and $\hat{K} = 46$ segments. With the segmentation only, the number of segments is estimated to be 70, which is significantly larger. Figures 4 and 3 give the four series with the estimated breakpoint positions (vertical lines) obtained by taken into account the dependency or not respectively. Accounting for the dependency is important to avoid too many false breakpoint detection (the additional breakpoints do not correspond to known events, see Table 2).

Covariance structure The choice $\hat{Q} = 1$ seems to be sufficient to capture the dependency structure among series. Σ well captures the spatial dependency between series as shown in Figure 5 which represents the estimation of Σ according to the distance between slave series (available here) and where we observed that the larger the distance, the higher the correlation.

Breakpoint interpretation Concerning the estimated breakpoint positions, we can distinguish:

- the common breakpoints among almost the four series (coded by the dotted lines (in red) in Figure 4). These breakpoints can be due to changes in the reference series MPLE. Among them, the breakpoint at day 54228 is known to be due to antenna and receiver changes (see Table 2).
- the series-specific changes. Some of them are due to known instrumental changes (the dashed lines (in black)). Moreover the change at day 53620 in the series 4 can be due to the known change in the series MPLE (see Table 2). The others (the solid lines (in black)) remain, up to now, unexplained even if some are close to known events.

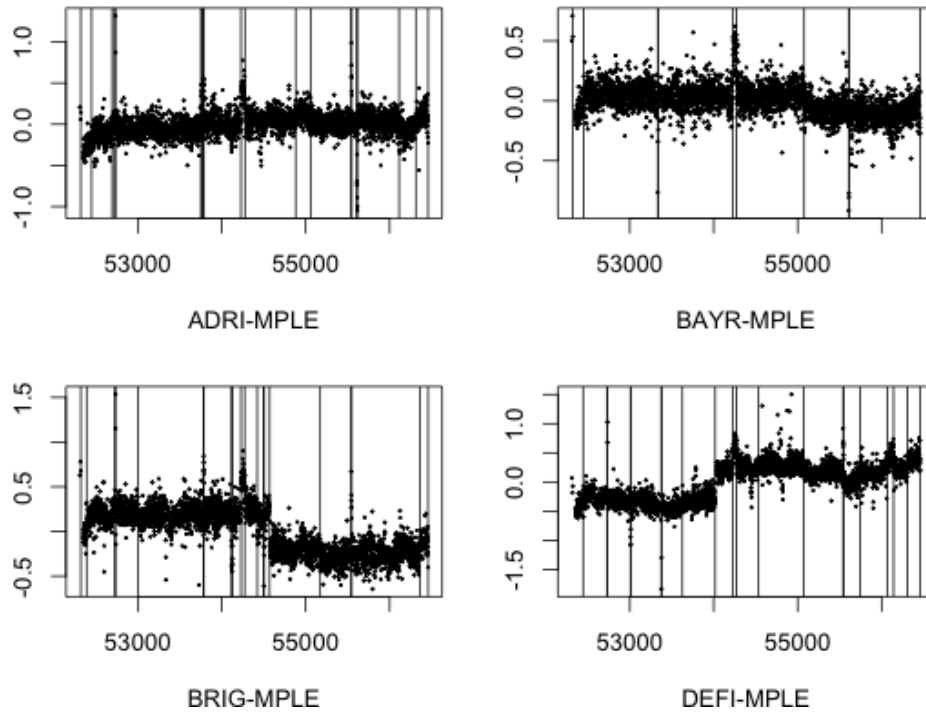


Figure 3: Estimated breakpoints for the 4 series obtained with the segmentation only.

We also observed that in the series 2 (BAYR-MPLE), the change at day 55077 is not referred as known event but seems well marked. Moreover, the series 4 (DEFI-MPLE) is the fastest series from MPLE. This can be explained by the periodic behavior observed in the series of difference (the tectonic motion part is not completely corrected) and so a large number of estimated breakpoints is needed to fit it.

References

- [1] J. Bai and P. Perron, *Computation and analysis of multiple structural change models*, J. Appl. Econ. **18** (2003), 1–22.
- [2] H. Caussinus and O. Mestre, *Detection and correction of artificial shifts in climate series*, JRSS-C **53** (2004), no. 3, 405–425.
- [3] C. Friguet, M. Kloareg, and D. Causeur, *A factor model approach to multiple testing under dependence*, J. Amer. Statist. Assoc. **488** (2009), 1406–15.
- [4] Julien Gazeaux, Simon Williams, Matt King, Machiel Bos, Rolf Dach, Manoj Deo, Angelyn W Moore, Luca Ostini, Elizabeth Petrie, Marco Roggero, Felix Norman Teferle, German Olivares, and Frank H. Webb, *Detecting offsets in gps time series: First results from the detection of offsets in gps experiment*, Journal of Geophysical Research: Solid Earth **118** (2013), no. 5, 2397–2407.

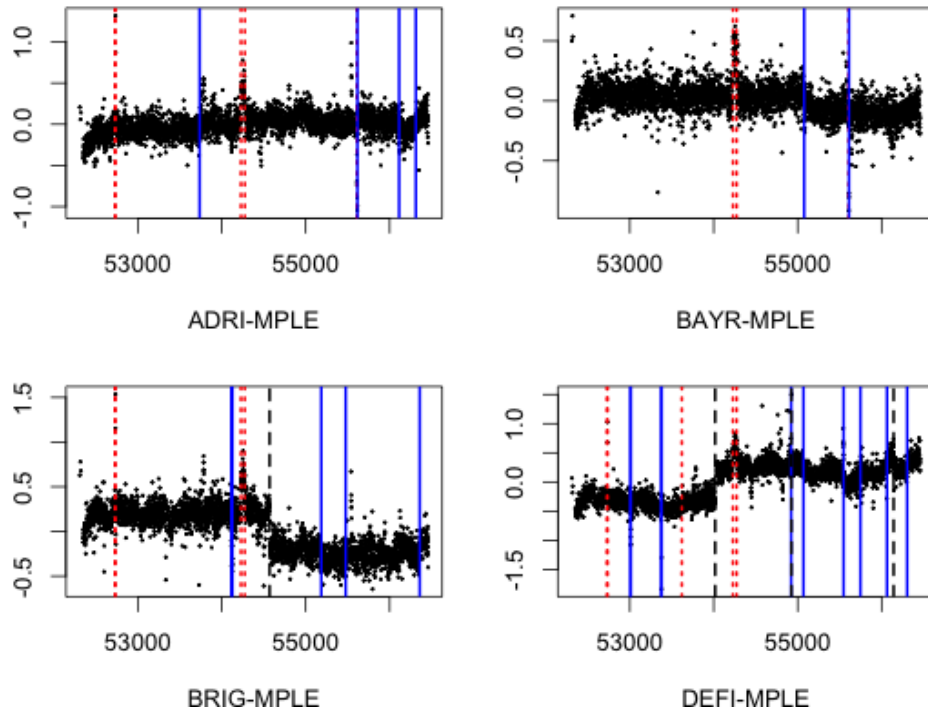


Figure 4: Estimated breakpoints for the 4 series when the dependency is taken into account. Dotted lines (in red): common breakpoints among almost the 4 series. Solid line (in blue): series-specific breakpoints. Dashed lines (in black): known series-specific breakpoints.

Serie	date	time in week	description of the change
ADRI	2004-12-02	53341	receiver
ADRI	2005-09-15	53628	receiver
ADRI	2006-08-02	53949	receiver
BAYR	2004-12-02	53341	receiver
BAYR	2005-08-31	53613	receiver
BAYR	2006-07-11	53927	receiver
BRIG	2004-12-02	53341	receiver
BRIG	2005-09-12	53625	receiver
BRIG	2008-04-14	54570	antenna
DEFI	2006-10-11	54019	antenna
DEFI	2009-04-07	54928	receiver
DEFI	2011-01-26	55587	receiver
DEFI	2012-08-02	56141	antenna
DEFI	2012-11-05	56236	antenna
MPLE	2004-12-02	53341	receiver
MPLE	2005-09-14	53627	receiver
MPLE	2006-12-28	54097	receiver
MPLE	2007-05-08	54228	antenna and receiver
MPLE	2007-10-05	54378	receiver

Table 2: Known changes in the five series.

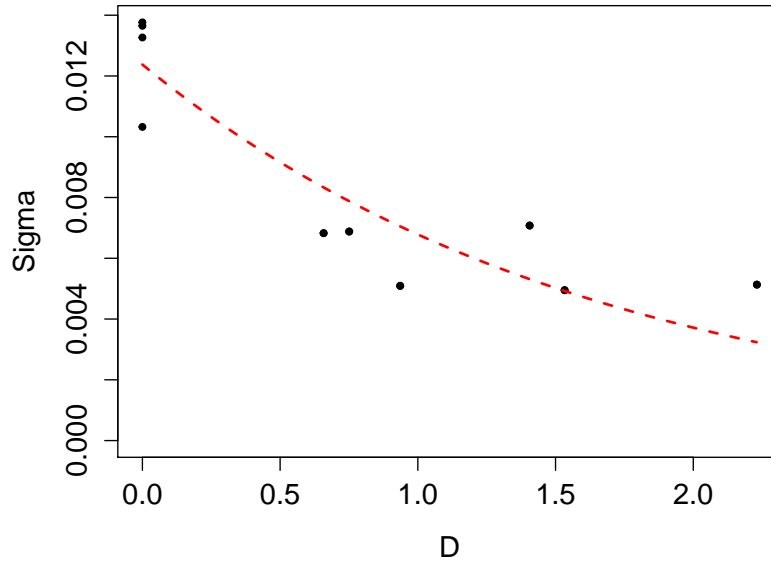


Figure 5: Estimation of Σ according to the distance between series.

- [5] W.R. Lai, M.D. Johnson, R. Kucherlapati, and P. J. Park, *Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data*, Bioinformatics **0** (2005), no. 0, 1–8.
- [6] O. Mestre, P. Domonkos, F. Picard, I. Auer, S. Robin, E. Lebarbier, R. Bhm, E. Aguilar, J. Guijarro, G. Vertachnik, M. Klancar, B. Dubuisson, and P. Stepanek, *Homer : a homogenization software - methods and applications*, Quarterly Journal of the Hungarian Meteorological Service **117** (2013), no. 1, 47–67.
- [7] F. Picard, E. Lebarbier, E. Budinska, and S. Robin, *Joint segmentation of multivariate gaussian processes using mixed linear models*, Comput. Statist. and Data Analysis (2011), no. 2, 1160–70.
- [8] F. Picard, E. Lebarbier, M. Hoebeke, G. Rigai, B. Thiam, and S. Robin, *Joint segmentation, calling and normalization of multiple cgh profiles*, Biostatistics **12** (2011), no. 3, 413–428.
- [9] F. Picard, S. Robin, M. Lavielle, C. Vaisse, and J.-J. Daudin, *A statistical approach for array CGH data analysis*, BMC Bioinformatics **6** (2005), no. 27, 1.
- [10] Donald B Rubin and Dorothy T Thayer, *EM algorithms for ml factor analysis*, Psychometrika **47** (1982), no. 1, 69–76.
- [11] D.A. van Dyk, *Fitting mixed-effects models using efficient EM-type algorithms*, Jour. Comp. and Graph. Statistics **9** (2000), 78–98.
- [12] S. Williams, *Offsets in global positioning system time series*, Journal of Geophysical Research: Solid Earth **108** (2003), 2310.
- [13] N. R. Zhang and D. O. Siegmund, *A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data*, Biometrics **63** (2007), no. 1, 22–32.